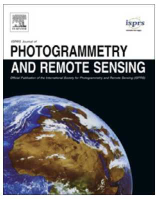
ISPRS Journal of Photogrammetry and Remote Sensing 122 (2016) 179–191

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs



A co-training, mutual learning approach towards mapping snow cover from multi-temporal high-spatial resolution satellite imagery

Liujun Zhu [a,b,c,d], Pengfeng Xiao [a,b,c,*], Xuezhi Feng [a,b,c], Xueliang Zhang [a,b,c], Yinyou Huang [a,b,c], Chengxi Li [a,b,c]

[a] *Department of Geographic Information Science, Nanjing University, Nanjing, Jiangsu 210023, China*
[b] *Collaborative Innovation Center of South China Sea Studies, Nanjing, Jiangsu 210023, China*
[c] *Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, Jiangsu 210023, China*
[d] *Department of Civil Engineering, Monash University, Clayton, Vic. 3800, Australia*

ARTICLE INFO

Article history:
Received 27 May 2016
Received in revised form 29 October 2016
Accepted 6 November 2016

Keywords:
Co-training
Multitask learning
Multi-temporal
Snow cover
Semi-supervised learning
High spatial resolution image

ABSTRACT

High-spatial and -temporal resolution snow cover maps for mountain areas are needed for hydrological applications and snow hazard monitoring. The Chinese GF-1 satellite is potential to provide such information with a spatial resolution of 8 m and a revisit of 4 days. The main challenge for the extraction of multi-temporal snow cover from high-spatial resolution images is that the observed spectral signature of snow and snow-free areas is non-stationary in both spatial and temporal domains. As a result, successful extraction requires adequate labelled samples for each image, which is difficult to be achieved. To solve this problem, a semi-supervised multi-temporal classification method for snow cover extraction (MSCE) is proposed. This method extends the co-training based algorithms from single image classification to multi-temporal ones. Multi-temporal images in MSCE are treated as different descriptions of the same land surface, and consequently, each pixel has multiple sets of features. Independent classifiers are trained on each feature set using a few labelled samples, and then, they are iteratively re-trained in a mutual learning way using a great number of unlabelled samples. The main principle behind MSCE is that the multi-temporal difference of land surface in spectral space can be the source of mutual learning inspired by the co-training paradigm, providing a new strategy to deal with multi-temporal image classification. The experimental findings of multi-temporal GF-1 images confirm the effectiveness of the proposed method.

1. Introduction

Snow cover extent is an essential input for snow hydrological models and snow hazard monitoring. Consequently, snow cover maps have been extracted using Moderate Resolution Imaging Spectroradiometer (MODIS) (Hall et al., 2002), Landsat Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+) (Crawford et al., 2013), Operational Landsat Imager (OLI) (Zhu et al., 2015) images, and ground based digital camera (Bernard et al., 2013). However, these sensors can unilaterally reach the high spatial or temporal resolution required to capture the seasonal spatial and temporal variations of snow cover. The new generation of satellites in the form of a constellation, such as Europe's Sentinel-2 and China's GF-1/6, can provide relatively high spatial resolution and frequently revisited observations, thereby providing spatial and temporal details of snow cover characteristics.

Current snow cover extraction approaches are mainly based on the special spectral characteristics of snow, i.e. high reflectance at visible wavelengths and low reflectance at shortwave-infrared wavelengths (Warren, 1982). A series of thresholds based on Normalized Difference Snow Index (NDSI) and/or spectral band ratio, are sufficient to separate snow from snow-free (Dozier, 1989; Hall et al., 1995; Riggs et al., 1994). Snow cover in mountain areas in shadow has a large overlap with the snow-free region in the spectral space (Dozier, 1989; Rosenthal and Dozier, 1996). Consequently, topographic correction based on the digital elevation model (DEM) is used to alleviate the influence of mountain shadow (Dozier, 1989; Negi et al., 2009; Rosenthal and Dozier, 1996; Sirguey et al., 2009). Alternatively, some studies apply shadow masks achieved from DEM to exclude the shadow areas (Selkowitz and Forster, 2016). To represent the subgrid snow cover heterogeneities, empirical relationships based on NDSI

[*] Corresponding author at: Department of Geographic Information Science, Nanjing University, China.
E-mail address: xiaopf@nju.edu.cn (P. Xiao).

http://dx.doi.org/10.1016/j.isprsjprs.2016.11.003
0924-2716/© 2016 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

(Salomonson and Appel, 2004, 2006) and subpixel unmixing models (Painter et al., 2009, 1998; Rosenthal and Dozier, 1996) are proposed to produce fractional snow cover maps. In addition, machining learning techniques, e.g. Artificial Neural Network and Support Vector Machine (SVM), are applied to train more robust models or classifiers for extracting both binary and fractional snow cover (Dobreva and Klein, 2011; Simpson and McIntire, 2001; Zhu et al., 2014).

Despite the great advance, it is still challenging to extract snow cover from high-spatial and -temporal resolution remote sensing images (HSTRRS). Commonly, shortwave-infrared wavelengths are not covered by high-spatial resolution optical sensors. As a result, subgrid snow cover heterogeneities become less critical and NDSI is not available. Some indexes based on visible and near-infrared bands provide promising alternatives for snow cover extraction in plains (Hinkler et al., 2003, 2002) but cannot obtain sound results in the mountain areas because of the severe influence of mountain shadow in HSTRRS. What's worse, high-spatial resolution DEM with satisfactory quality is not commonly accessible, making topographic correction difficult.

In our previous study (Zhu et al., 2014), a SVM based decision tree was proposed to extract snow cover from a single high-spatial resolution image without topographic correction, where snow cover influenced by mountain shadow was treated as an independent class in the classification procedure. Similar to most of the classification methods, the main limitation of this method is its heavy dependency on the quality of the ground-truth samples. Collecting a sufficient number of representative samples is impractical. Moreover, even if a satisfactory classifier was trained for an image with adequate samples, it cannot be directly applied to other acquisitions, because the observed spectral distributions of different images can be different for many reasons, e.g. variations in the observation geometry and mountain shadow. Therefore, a more robust method without heavy dependency on labelled samples and cumbersome topographic correction is needed.

Domain adaptation (also known as transfer learning) is one of the most promising methods to solve this problem. In the domain adaptation paradigm, a strong classifier is trained for a specific image (source domain) with adequate labelled samples and this classifier is then applied to a new acquisition (target domain) with the assistance of unlabelled samples (Liu and Li, 2014). This kind of method has been used in remote sensing classifications (Kurtz et al., 2014; Liu and Li, 2014), especially for the automatic updating of land cover maps (Bahirat et al., 2012; Bruzzone and Marconcini, 2009; Matasci et al., 2015). The main challenge for the application of this method to extract snow cover maps may be that it still needs sufficient labelled samples for a specific image or at least a part of an image.

Semi-supervised learning is another kind of promising approach, which can use a few labelled samples together with unlabelled samples to increase the reliability and accuracy of a classifier. Four paradigms of semi-supervised learning are encountered in literature, i.e. generative models (Shahshahani and Landgrebe, 1994), low density separation algorithms (Joachims, 1999; Vapnik, 1998), graph-based methods (Jordan, 1998), and co-training algorithms (Blum and Mitchell, 1998). All these semi-supervised methods have been successfully applied in remote sensing classifications with a small group of labelled samples (Bruzzone et al., 2006; Camps-Valls et al., 2007; Dalponte et al., 2015; Jackson and Landgrebe, 2001; Tan et al., 2014). However, these methods are merely suitable for the classification of single images and needs to be extended to deal with the multi-temporal classification.

In this study, we proposed a strategy to extend aforementioned co-training (also known as multi-view learning) algorithms from single image classification to multi-temporal ones so that a few labelled samples are sufficient to extract snow cover maps from multi-temporal images simultaneously. In the original concept of co-training, the feature set (e.g. spectral or texture features in terms of remote sensing data) should be split into two subsets, where each subset should be sufficient for training a strong classifier, and these classifiers are conditionally independent of each other for a given class label (Blum and Mitchell, 1998). The process of co-training is rather simple. Two classifiers are trained on two subsets for the same task first. Then these classifiers provide each other with labels for the unlabelled data. The unlabelled samples here serve as a "platform" for information exchange (Zhou and Li, 2010). Further studies showed that the assumption of two conditionally independent feature subsets was not necessary (Wang and Zhou, 2007). The key for the co-training approaches to succeed is that there exists a large difference between the classifiers, while it is not crucial in which the difference is introduced (Dasgupta et al., 2002; Wang and Zhou, 2007). Many variants of co-training have been proposed, e.g. an improved algorithm combining co-training with Expectation-Maximization (Co-EM) (Nigam and Ghani, 2000) and a further integration with SVM (Co-EM-SVM) (Brefeld and Scheffer, 2004).

For the snow cover extraction, the multi-temporal images provide multiple descriptions (multiple feature subsets) of the same snow cover area and these feature sets of snow cover can be different for many reasons, e.g. ageing of snow, contamination caused by dust, change of illumination, and observation geometry. As a result, the classifiers respectively trained on different images have a large difference, and the mutual learning based on the difference can, therefore, be used to improve the reliability of classifiers. Fig. 1 depicts the relationship between the conception of original co-training methods and the multi-temporal extension one for snow cover extraction from HSTRRS. It is expected that a few labelled samples are sufficient to extract snow cover from HSTRRS simultaneously by using the multi-temporal extension of co-training.

However, several issues should be carefully considered in the use of the multi-temporal extension of co-training methods. While different feature subsets split from one feature set naturally have the same labels in the co-training methods, it is not true in multi-temporal cases because of the possible transition between the snow and snow-free areas in different acquisitions. In addition, there are a large number of unlabelled samples in remote sensing. A selection procedure is necessary to choose proper unlabelled samples that can enhance the mutual learning. Furthermore, the co-training methods are inherent two-class methods. Further extension is needed to deal with the multi-temporal multi-class problems. In this study, these issues are addressed to extend the Co-EM-SVM from a single image classification to a multi-temporal one. It is worth noting that other co-training algorithms can also be extended to multi-temporal methods in a similar way.

The rest of this paper is organized in six sections. A brief introduction to Co-EM-SVM, followed by the proposed method, is presented in Section 2. The study area and data are described in Section 3. In Section 4, the experimental design is introduced. The performance of the proposed method is evaluated in Section 5. Sections 6 and 7 are discussions and conclusions, respectively.

2. Methodology

2.1. Co-EM-SVM

Co-EM-SVM (Brefeld and Scheffer, 2004) is an improved variant of co-training (Blum and Mitchell, 1998) and Co-EM (Nigam and Ghani, 2000). In Co-EM-SVM, the available feature set V of a data set is split into disjoint sets V_1 and V_2 (V_i is a collection of some

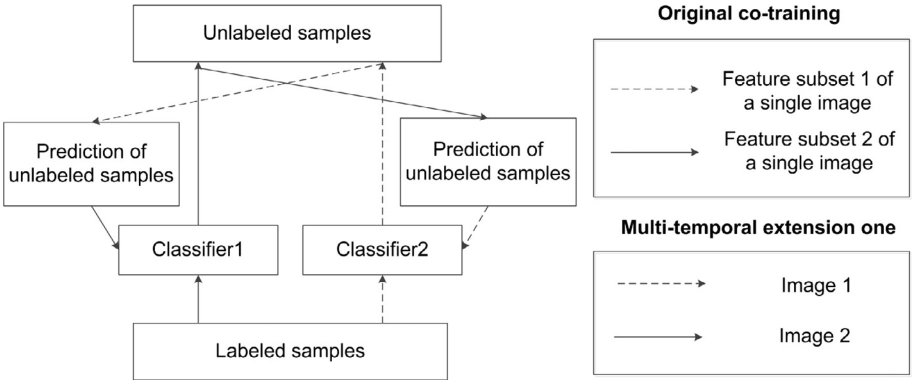

Unlabeled samples
Original co-training
Feature subset 1 of a single image
Feature subset 2 of a single image
Prediction of unlabeled samples
Prediction of unlabeled samples
Classifier1
Classifier2
Multi-temporal extension one
Image 1
Image 2
Labeled samples

Fig. 1. Relationship between the conception of original co-training algorithm and the multi-temporal extension one for snow cover mapping from HSTRRS. Two feature subsets of a single image required in the original co-training algorithm are replaced by two images acquired over the same location in the multi-temporal extension one.

spectral bands in terms of remote sensing data). That is, each sample (x, y) of the labelled sample set D_l is decomposed and viewed as (x^1, x^2, y), where x^1 and x^2 are the vectors over V_1 and V_2, respectively, and $y \in \{1, -1\}$ is the label of the sample. The procedure of Co-EM-SVM is similar to co-training (Fig. 1). Two SVMs f^1 and f^2 are trained on V_1 and V_2 using a few labelled samples, respectively. These SVMs then estimate the class probabilities $p(y|x^{1*})$ and $p(y|x^{2*})$ of the unlabelled sample set $D_u(x^{1*}, x^{2*})$. Labelled and unlabelled samples together with class probabilities are used to re-train the classifiers iteratively. The detail of Co-EM-SVM can be found in Brefeld and Scheffer (2004) and two key steps are described below.

Step 1: Calculating $p(y|x)$ for unlabelled samples. The prior probabilities $p(y)$ are calculated from the labelled data. Subsequently, the decision values of SVM for a class $p(f(x)|y)$ are assumed to be governed by a normal distribution $N[\mu, \sigma^2]$. μ_y and σ^2 are calculated as

$$\mu_y = \left(\sum_{D_l^y} f(x) + \sum_{D_u^y} f(x)\right) \Big/ (N_l + N_u), \tag{1}$$

$$\sigma_y^2 = \left(\sum_{D_l^y \cup D_u^y} (f(x) - \mu_y)^2\right) \Big/ (N_l + N_u), \tag{2}$$

where D_u^y is a subset of D_u, whose label is predicted as y, N_l and N_u are the number of labelled and unlabelled samples for each class, respectively. Then $p(y|x)$ $(x \in D_u)$ can be calculated by

$$p(y|x) = \frac{N[\mu_y, \sigma_y^2](f(x))p(y)}{\sum_i N[\mu_y, \sigma_y^2](f(x))p(y)}, \tag{3}$$

where $y \in \{1, -1\}$.

Step 2: Training SVMs using D_u and D_l. The aim is to train a hyperplane $h : f(x) = w \cdot x + b = 0$ on each feature subset, where w is a vector (normal to h), and b is the bias (a constant). This hyperplane can be trained by solving the optimization problem:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}|w|^2 + C\sum_{j=1}^{N_l} \xi_j + C_s \sum_{j=1}^{N_u} c_{x_j^*} \xi_j^*, \tag{4}$$

$$\text{subject to } \forall_{j=1}^{N_l} y_j(wx_j + b) \geqslant 1 - \xi_j, \ \forall_{j=1}^{N_u} y_j^*(wx_j^* + b) \geqslant 1 - \xi_j^*, \tag{5}$$

$$\forall_{j=1}^{N_l} \xi_j > 0, \quad \forall_{j=1}^{N_u} \xi_j^* > 0, \tag{6}$$

where ξ_j and ξ_j^* are the slack variables for labelled and unlabelled samples, respectively. C and C_s are the regularization parameters for the labelled and unlabelled samples, respectively. For an unlabelled sample x_j^* with a predicted label $y_j^* = \arg\max_y p(y|x_j^*)$, its weight is defined as $c_{x_j^*} = p(y = y_j^*)(\max p(y|x_j^*) - \min p(y|x_j^*))$.

Four parameters are required for Co-EM-SVM, namely, the regularization parameters C and C_s, the number of iterations *iter*, and the prior probabilities $p(y)$ of the unlabelled samples. C and *iter* can be provided by a parameter selection procedure or defined by the users. C_s is set as a small value initially and then is doubled after each round, which is designed to reduce the risk of finding the local minima (Joachims, 1999). The initial C_s is set to $C/2^{(iter+1)}$ in this study. A selection procedure for the unlabelled samples, which can provide an accurate estimation of $p(y)$, will be introduced in the following subsection.

Similar to the original co-training, if the classifiers trained on two feature sets have a large difference and each feature set is sufficient to train a strong classifier (Brefeld and Scheffer, 2004), Co-EM-SVM can improve the performance of the classifiers.

2.2. Multi-temporal extension of Co-EM-SVM

2.2.1. Framework of the proposed MSCE

The main conception of the multi-temporal extension of Co-EM-SVM is shown in Fig. 1. Two feature subsets of a single image required in the Co-EM-SVM are replaced by two images acquired over the same location. A more straightforward description is presented in Fig. 2. The input includes two images T_1 and T_2 achieved successively over the same area with a few labelled samples, which are denoted by the red and blue points in Fig. 2(a). In the first step, some unlabelled samples (green points) are selected automatically. The details of generating satisfactory unlabelled samples will be introduced later. Note that the labelled and unlabelled samples have consistent labels in images T_1 and T_2, although the identity of the unlabelled sample is unknown. In general, these samples have different distributions in the feature space shown in Fig. 2 (b), which is the key for the proposed method to succeed. In the second step, the labelled samples are used to train a pair of initial classifiers shown as the black line in Fig. 2(c), and then these classifiers are used to estimate the class probabilities of the unlabelled samples. In the third step as shown in Fig. 2(d), the unlabelled samples with class probabilities given by one classifier are used to retrain the other classifier. The second and third steps are executed iteratively until a predefined iteration number is reached or the two classifiers yield the same prediction on the unlabelled samples. Note that these two classifiers can give different predictions on other samples as they have quite different hyperplanes. Finally, the trained classifiers are used to extract snow cover respectively. Similar to the original Co-EM-SVM, if the classifiers trained for two

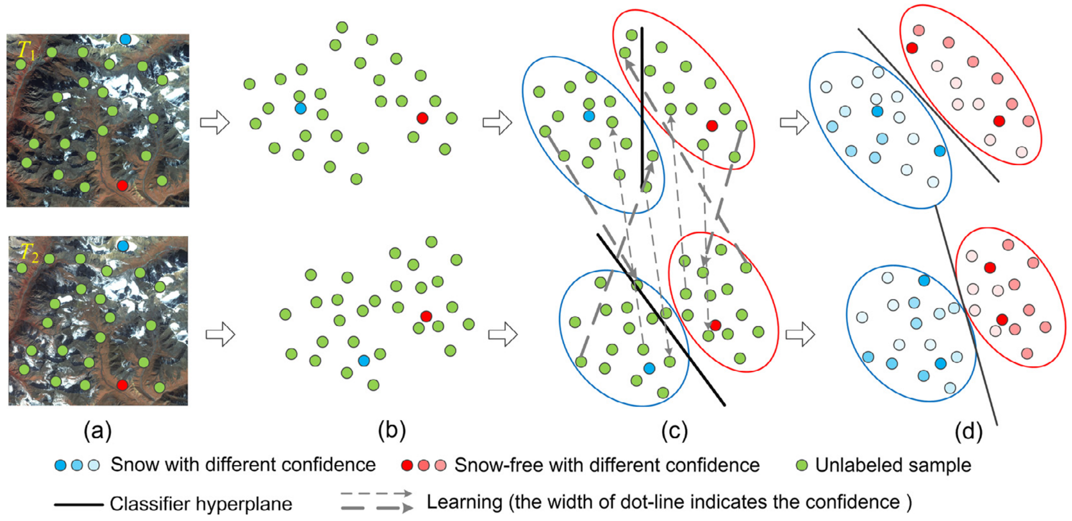

T_1
T_2
(a)
(b)
(c)
(d)
●●○ Snow with different confidence
●●○ Snow-free with different confidence
● Unlabeled sample
Classifier hyperplane
Learning (the width of dot-line indicates the confidence)

Fig. 2. Mutual learning procedure of MSCE. (a) Two images acquired on different phases, together with the samples (two circles with the same location in two images correspond to a sample); (b) distributions of samples in feature space (distributions are similar, but not identical); (c) learning process where the unlabelled samples with different confidence values are used to teach the other classifier (green circles linked with dashed line belong to the same sample); (d) learning result together with the re-trained classification hyperplanes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

images have a large difference and each image is sufficient to train a strong classifier, this method is expected to train two satisfied classifiers.

2.2.2. Selection of unlabelled samples for training

As mentioned in the introduction, a sample has the same label in two feature subsets in the original co-training paradigm. This is not true for some parts of the multi-temporal images as a result of snow melting or snowfalls. A pre-selection procedure of change detection can find the unlabelled samples with constant labels in the multi-temporal images. In this study, an unsupervised change detection algorithm is selected because of its independence from the labelled samples.

Specifically, a difference image is first calculated by Chi-Square Distance (CSD) (D'Addabbo et al., 2004). Given two images I_1 and I_2, x_i^1 and x_i^2 are the ith pixels of I_1 and I_2, respectively. The CSD of the ith sample (x_i^1, x_i^2) is computed by

$$CSD = \sum_{k=1}^{N}\left(\frac{x_{i,k}^1 - x_{i,k}^2}{\sigma_{diff}^k}\right)^2, \qquad (7)$$

where N is the number of bands and σ_{diff}^k is the standard deviation of the kth spectral band in the difference image. Then, the unsupervised Kittler–Illingworth thresholding algorithm (Kittler and Illingworth, 1986) is used to automatically classify the difference image into two classes, namely, the changed and unchanged areas. Pixels located in the unchanged area are chosen as the candidate unlabelled sample set $D_u(x^{1*}, x^{2*})$. It is worth noting that adequate unlabelled samples can always be found as the number of unlabelled samples required is much smaller than the huge amount of pixels in remote sensing. Moreover, the changed area commonly accounts for a small part of the image.

A further selection procedure is proposed based on the characteristics of Co-EM-SVM and the fact that the candidate unlabelled sample set is still too large for training. In the Co-EM-SVM, there is no selection for the unlabelled samples and the key parameter $p(y)$ of the unlabelled samples is assigned identically to that of the labelled samples. The discrepancy between the $p(y)$ of unlabelled and labelled samples can introduces large uncertainty, especially when only few labelled samples are available. A further selection can help determine $p(y)$. In addition, each round of Co-EM-SVM requires a time equivalent to that of the standard SVM. This time will slightly decrease by reducing the number of unlabelled samples for each class (N_u). A relatively small N_u will speed up the learning process. Moreover, it is quite possible that one classifier will receive labels on the unlabelled samples erroneously weighted by the other classifier during the first round when the initial classifiers have moderate accuracy (Zhang and Zhou, 2011). Hence, this selection procedure should have the capability to select a set of "qualified" unlabelled samples for the first round of learning.

A "qualified" unlabelled sample set means: (i) *confidence*, samples have high-confidence initial labels, so that one classifier can use these samples to teach the other without harming it; (ii) *difference*, the selected unlabelled samples should inherit the difference contained in the original unlabelled sample set. However, these two points are conflicted because the unlabelled samples that have a high confidence for both classifiers are partially selected and cannot fully inherit the difference. Hence, a *confidence-difference* tradeoff selection procedure is proposed. Table 1 presents the procedure of selecting unlabelled samples for a binary problem, which includes the two steps given as follows.

Step 1: Selecting relatively confident samples. The unlabelled samples near the classification hyperplane have low confident labels than those with a larger distance. For a tradeoff between *confidence* and *difference*, a threshold criterion is employed to achieve the relatively confident samples without discarding too many unlabelled samples in this step, which may have a large difference over the two feature subsets. The threshold is defined as the product of the *confidence-difference* tradeoff parameter λ and the average predicted values $f(x)$ of each class. The unlabelled samples whose predicted values $f(x)$ are greater than or equal to the threshold are retained. When λ is set to a small value, the unlabelled samples have a low average confidence, but inherit a large difference. Since each unlabelled sample has two feature vectors, only the samples having a high predicted value to predict the same label for both images are retained. Fig. 3 illustrates why selecting the unlabelled samples with high predicted values can alleviate the risk of "bad" learning.

Step 2: Selecting samples from the relatively confident samples. In most cases, the number of relatively confident samples is

Table 1
Procedure of selecting unlabelled samples for a binary problem.

```
Input:
Labelled sample set $D_l(x^1, x^2, y)$, candidate unlabelled sample set $D_u(x^{1*}, x^{2*})$,
    number of unlabelled samples $N_u$ to be selected for each class, and
    threshold parameter $\lambda$.
Begin:
  For i = 1, 2
  • Train a SVM $f^i$ on $(x^i, y)$
  • Compute the decision values $f^i(x^{i*})$ of unlabelled samples and classify
    them into positive and negative classes $\Psi_i^{\pm}$ according to the predict label
    $y^i$
  • Compute the positive and negative threshold values $Th_i^{\pm}$:
      $Th_i^+ = \lambda \times \mathrm{mean}(f^i(x_p^{i*})), x_p^{i*} \in \Psi_i^+$
      $Th_i^- = \lambda \times \mathrm{mean}(f^i(x_n^{i*})), x_n^{i*} \in \Psi_i^-$
  End
  • Achieve the unlabelled samples with high confidence
      $D_u^+ = \{(x_p^{1*}, x_p^{2*}) | x_p^{i*} \in \Psi_i^+, f^i(x_p^{i*}) \geqslant Th_i^+\}$
      $D_u^- = \{(x_n^{1*}, x_n^{2*}) | x_n^{i*} \in \Psi_i^-, f^i(x_n^{i*}) \leqslant Th_i^-\}$
  • Randomly select $N_u$ samples from $D_u^+$ and $D_u^-$ respectively to define $D_u$
End
Output: $D_u(x_j^{1*}, x_j^{2*}),\quad j = 1, \ldots, 2N_u$
```

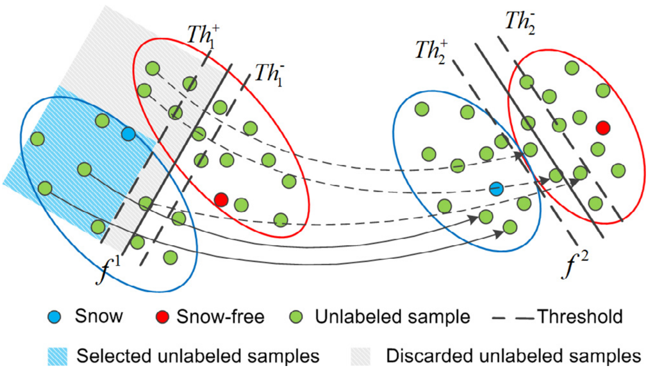


Fig. 3. Example of selecting confident unlabelled samples. The unlabelled samples linked by the dashed line are samples highly confident to f^1, while they are harmful information for f^2. These samples will evidently harm the re-training f^2. Only the unlabelled samples with high confidence for both classifiers are selected, which at least do not contain wrong information.

very large. To further reduce the number of selected unlabelled samples and determine $p(y)$, a fixed number (N_u) of the unlabelled samples for each class are randomly selected. After this procedure, the unlabelled sample set $D_u(x_j^{1*}, x_j^{2*}), j = 1, \ldots, 2N_u$, together with the labelled sample set $D_l(x^1, x^2, y)$ are treated as the input of binary Co-EM-SVM. Since the number of selected samples for each class has the same value of N_u, the prior probabilities $p(y)$ required in Co-EM-SVM is a fixed value of 0.5.

2.2.3. Multiclass solution

Let $\Omega = \{\omega_k\}_{k=1}^{M}$ be a set of M possible classes. There are several multiclass strategies for binary SVMs, including one against all (OAA), one against one (OAO), and directed acyclic graph (DAG) (Hsu and Lin, 2002). However, only the OAA strategy can be employed in the proposed method and the reasons were well established in Bruzzone et al. (2006). Briefly, in the selection of unlabelled samples and learning process of Co-EM-SVM, each sample must belong to one of the two classes Ω^- or Ω^- ($\Omega^- \cup \Omega^+ = \Omega$). This means that all the unlabelled samples should be classified into two classes. Thus, the strategies using pairwise classifiers, i.e. OAO and DAG, cannot work (Bruzzone et al., 2006).

Based on OAA strategy, an ensemble of M (parallel) binary classification problems $\{\omega_k, \Omega - \omega_k\}$, $k = 1, \ldots, M$ is built. Then the learning procedure of Co-SVM-EM on each problem is carried out. Two sets of SVM classifiers $\{f_1^1, \ldots, f_k^1, \ldots, f_M^1\}$ and $\{f_1^2, \ldots, f_k^2, \ldots, f_M^2\}$ are trained for I_1 and I_2, respectively, as shown in Fig. 4. Finally, the original images I_1 and I_2 are input to the corresponding classifier sets and a "winner-takes-all" rule is used to decide the label of each sample, which can be expressed as $\omega = \arg\max_k\{f_k^i(x^i)\}$, as shown in Fig. 4.

3. Study area and data

3.1. Study area

The study area is located in the southwest of the Manasi River Basin on the north submontane of Tianshan Mountains, Xinjiang Province, China. This region is in an alpine area with high topographic relief and complex terrain conditions (Fig. 5). Specifically, the elevation of this region ranges from 2549 m to 4447 m. The proportion of area where the slope is greater than 30° is 29.4% and the influence of mountain shadows is thus severe. The main part of the study area is seasonally covered in snow, where the snow accumulation period extends from November to February of following year. Mountain belts higher than 3900 m (permanent snow line) are covered with snow and glaciers (Cheng et al., 2006).

3.2. Data

As the first satellite of the China's High-resolution Earth Observation System, GaoFen-1 (GF-1, which means high resolution in Chinese) was launched on April 26, 2013. There are four wide-field-of-view sensors (WFV) and two panchromatic and multispectral sensors (PMS) aboard on GF-1. The PMS has four spectral bands with a spatial resolution of 8 m, i.e. B1 (0.45–0.52 μm), B2 (0.52–0.59 μm), B3 (0.63–0.69 μm), and B4 (0.77–0.89 μm). The side swing capability of GF-1 allows PMS a short revisit period of 4 days, which can be further enhanced after the launch of GF-6 (a similar satellite of GF-1 scheduled for 2016). The spatial error is less than 50 m without ground control points (Bai, 2013). Part of the GF-1 data is freely available at http://www.cresda.com/EN/.

Three PMS images (T1, T2, and T3) acquired on October 7, 15, and 19 in 2013 were used in the study, which are depicted in Fig. 6. There was a snowfall between T1 and T2, followed by a snow melting process which can partly reflect the snow accumulation and melting processes in a snow season. According to the pre-

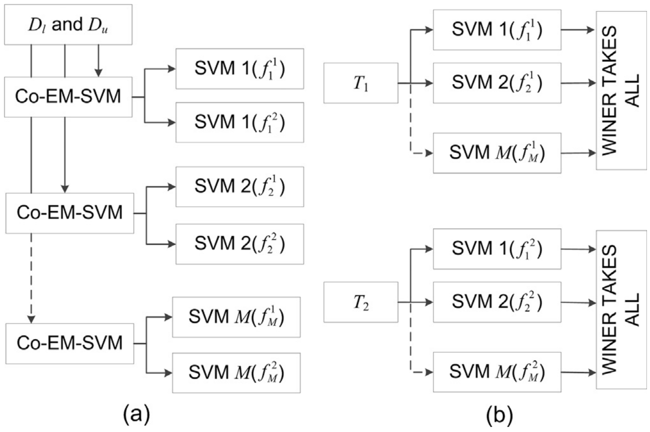


Fig. 4. Multiclass solution for the proposed method. (a) and (b) represent the training and classification procedure, respectively.

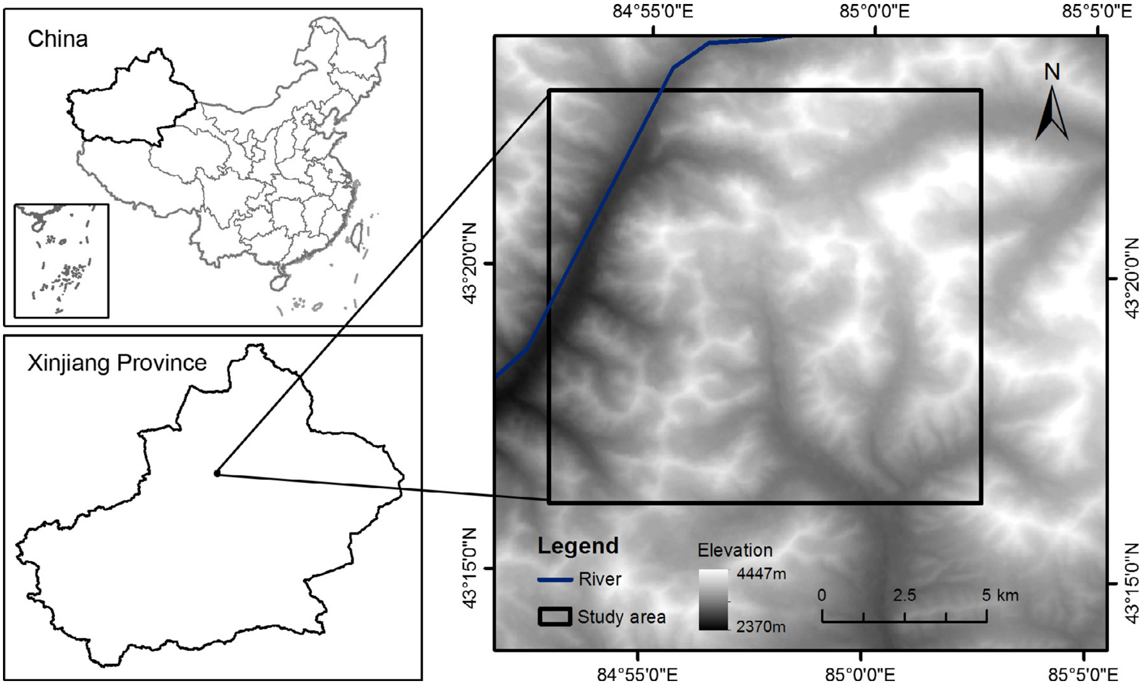


Fig. 5. Illustration of the study area.

launch radiometric calibration parameters, the visible bands of PMS are saturated when the top of atmosphere (TOA) reflectance is greater than 0.75 during the acquisitions. As a result, the fresh snow in sunlight (reflectance > 0.9) is saturated at these bands. Fortunately, the ageing of snow can still be captured by the near-infrared band. In addition, the snow in shadow has much lower TOA and thus can be fully recorded.

In our previous study (Zhu et al., 2014), the classification strategy of considering the snow-covered region influenced by mountain shadows as an independent class achieves satisfactory accuracy in extracting snow cover in the mountain shadows. This strategy has also been used in this study. Hence, three classes were selected, namely, snow in sunlight, snow in shadow, and snow-free. Since no field experiment is carried out during the acquisitions, samples for training and validation were selected randomly from images and identified by visual interpretation. To select reliable training and validation samples, a DEM product (Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Digital Elevation Model, ASTER GDEM, 30 m resolution) was used as assistance. Specifically, hill-shade images and permanent snow cover areas (>3900 m) were determined using the DEM first. This process does not need to delineate accurate shadows and thus the quality and spatial resolution of DEM is not critical. We selected 600 snow samples from permanently snow-covered regions to form the training set. Half of them were located in the area without direct solar radiance. The same number of snow-free samples was selected from the regions that are not permanently covered by snow. Each training sample has an identical label on three images, which is required for MSCE. The validation sets were built respectively for the images. The detailed information of the training and validation sets is shown in Table 2.

4. Experimental design

4.1. Validation metric

F score was selected as the accuracy metric, which is defined as (Olson and Delen, 2008):

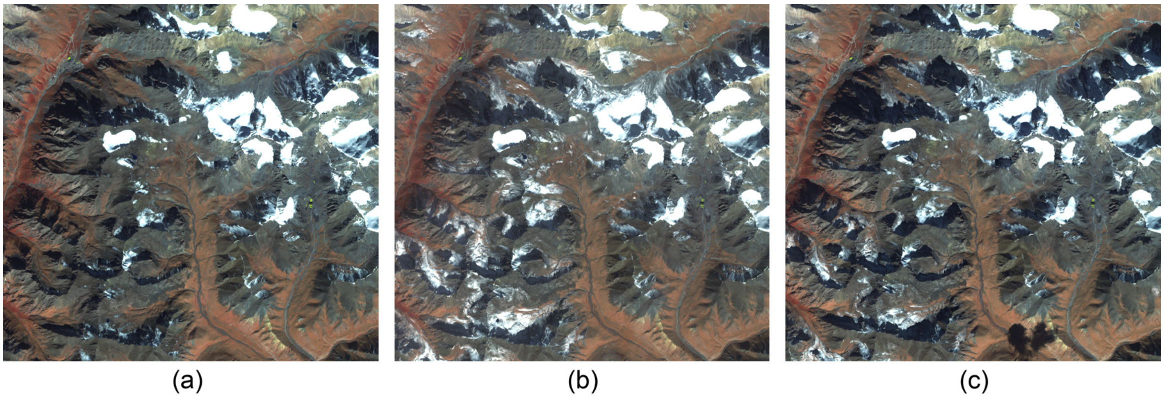


Fig. 6. (a), (b), and (c) are the images T1, T2, and T3 with the false-color composite (band 4, 3, and 2, respectively).

$$F = \frac{2TP}{2TP + FP + FN}, \tag{8}$$

where *TP* (true positive) is the number of snow pixels that are correctly identified, *FP* (false positive) is the number of snow-free pixels that are identified as snow, and *FN* (false negative) is the number of snow pixels that are identified as snow-free. The *F* score is treated as one of the most useful metrics because it penalizes both missing snow and falsely identified bare areas as snow (Rittger et al., 2013).

4.2. Experimental setup

Three experiments were designed to evaluate the performance of MSCE as well as the effect of unlabelled sample selection and spatial registration error. A summary of parameters required in MSCE is included in Table 3. Given a set of input parameters (σ, *C*, *iter*, T_i&T_j, N_l, N_u, and λ), snow cover were extracted 100 times, each of which is based on labelled samples randomly selected from the training set and evaluated using the validation set (Table 2). Two statistics of 100 validation results were calculated: mean and standard deviation (SD), which reveal extraction accuracy and stability, respectively. For simplicity, *F* score is used to denote the average *F* score of 100 trials hereafter.

The kernel of Radial Basis Function was used. The values of σ and *C* were selected on the basis of the training error using a coarse grid-searching method (Bruzzone et al., 2006). The candidate values for the σ were {0.125, 0.25, 0.5, 1, 2, 4, 8, 16} and those for *C* were {0.001, 0.01, 0.1, 1, 10, 100, 1000, 1000}. *iter* was 8 which is the same as in Brefeld and Scheffer (2004). Experiment-specific setup for N_l, N_u, λ and temporal combination is included in Table 4 and the purpose of each experiment is introduced as follows:

A. The first experiment is designed to evaluate the performance of MSCE comprehensively. N_l was set to small values (from 1 to 29 with an increment of 2) to show the performance of MSCE when limited labelled samples are available. All possible temporal combinations were included. N_u and λ was 100 and 1, respectively. The representative supervised and semi-supervised single-image classification method, i.e. SVM (Vapnik, 1998) and Transductive SVM (TSVM) (Bruzzone et al., 2006), were used for comparison. Other semi-supervised multi-temporal classification methods, e.g. domain adaptation methods, are not designed for the cases where only a few labelled samples are available. They cannot be directly compared with the proposed method and thus were not included.
The effect of temporal combination on accuracy was also investigated in this experiment. The performance of a combination is expected to associate with the two conditions of MSCE: (i) each image is sufficient to train a powerful classifier, and (ii) the classifiers trained on two images have a large difference. A powerful classifier naturally results in higher accuracy and vice versa. Thus, the validation results can be an indicator of the former. A simple metric is introduced to represent the prediction difference of the classifiers (PDC):

$$PDC(f^1, f^2) = 1 - \sum_{i=1}^{N_u} \delta(f^1(x_i^{1*}) - f^2(x_i^{2*}))/N_u, \tag{9}$$

where f^1 and f^2 denote two classifiers. (x_i^{1*}, x_i^{2*}) is an unlabelled sample. $\delta(x)$ is the Dirac Delta function which equals 1 if $x = 0$. When f^1 and f^2 have different prediction on all samples, PDC has a maximum value of 1.

B. The influence of two parameters related to the selection of unlabelled samples is investigated in the second experiment, which includes two sub-experiments. N_u was changed from a small value to a larger one to exploit the impact of N_u (Table 4). The temporal combination and λ were set to T1&T2 and 1, respectively. Since the optimal N_l is expected to tightly relate to N_u (Castelli and Cover, 1996), N_l was also changed from 1 to 29 with a transition at 2 to show its interactive effect with N_u on accuracy. Similarly, λ was changed from 0 to 1.8 with other parameters fixed to exploit its impact on accuracy in the second sub-experiment.

C. The spatial correspondence of multi-temporal image is one of the essential bases for the proposed method. Thus, the last experiment is designed to test the influence of spatial registration error. Specifically, we moved T2 in four directions with different pixels, i.e. north-south, northwest-southeast, west-east, and southwest-northeast, to introduce the spatial registration error manually. Then, the new T2 was combined with T1 to repeat the experiments.

Table 2
Information of classes and sample sets.

Class	Training set	Validation set		
		T1	T2	T3
Snow in sunlight	300	500	500	500
Snow in shadow	300	500	500	500
Snow-free	600	1000	1000	1000
Total	1200	2000	2000	2000

Table 3
A summary of parameters required in MSCE.

Parameter	Meaning
σ	Kernel width of Radial Basis Function
C	Regularization parameter
Iter	Iteration number of multi-learning
T_i&T_j	Temporal combination of image T_i and image T_j
N_l	Number of labelled samples for each class
N_u	Number of unlabelled samples for each class
λ	*Confidence-difference* tradeoff parameter

5. Results

5.1. Performance of MSCE

Fig. 7(a) and (d) shows the results of T1. In general, the *F* scores of three algorithms gradually increase when N_l increases. The *F* scores of MSCE changes from 0.874 to 0.905 (using the combination of T1&T2) and from 0.874 to 0.899 (using the combination of T1&T3), which is greater than that of SVM and TSVM in all cases. Specially, semi-supervised methods (MSCE and TSVM) have significantly better performance than SVM when only a sparse labelled set is available but this gap decreases when N_l increases. This can be explained by that, when more labelled samples are available, the real distribution of test data has been well represented by the labelled samples, resulting in a diminished effect of the unlabelled samples in defining the classification hyperplane. In addition, three algorithms reach high accuracies when N_l is greater than a certain number, which is 15 for SVM and 5 for MSCE and TSVM. This indicates that the unlabelled samples can help represent the general classification problem with fewer labelled samples. With respect to the SDs, MSCE has smaller SDs than SVM and TSVM in most cases. Particularly, the SDs of MSCE are much smaller than those of SVM and TSVM when only one sample is available for each class. The SDs of MSCE in the case $N_l = 1$ are 0.055 (T1&T2) and 0.061 (T1&T3), while these are doubled in the

Table 4
Details of the parameters for each experiment. A:B:C means the value of this parameter changes from A to C with a transition at B.

Experiment name	N_l	N_u	λ	Temporal combination
General performance	1:2:29	100	1	T1&T2, T1&T3, T2&T3
Selection of parameter N_u	1:2:29	10:20:90; 100:200:900	1	T1&T2
Selection of parameter λ	1:2:29	100	0:0.2:1.8	T1&T2
Influence of spatial registration error	1	100	1	T1&T2

results of TSVM (0.128) and SVM (0.104). In the results of T2 and T3, a similar phenomenon can be observed.

A comparison on different temporal combinations for each image is depicted in Fig. 8. For T1, the combination of T1&T2 achieves better results than T1&T3 owing to higher *F* scores and smaller SDs in all cases. The optimal combination for T2 and T3 is T1&T2 and T2&T3, respectively. This can be partly explained by the fact that a powerful "partner" can provide accurate label information, and thus, the other classifier can naturally benefit from mutual learning. Specifically, three algorithms achieve the best results in T2 followed by T1 and T3 (Fig. 7). This indicates that the classifier trained on T2 can provide more accurate label information than others. As a result, T1 has higher *F* scores in T1&T2 than in T1&T3 in all cases and T3 has higher accuracy in the combination of T2&T3 than in T1&T3.

To further illustrate the difference caused by temporal combinations, the PDC values of three temporal combinations before and after mutual learning were calculated using Eq. (9) and depicted in Fig. 9(a). In general, the initial classifiers have higher PDC values than the classifiers after learning. This is because, in a mutual learning process, the difference between two classifiers is utilized to improve the performance of the initial classifiers. The classifiers will gradually provide more consistent prediction for the same unlabelled samples, thereby decreasing the PDC. Further results are depicted in Fig. 9(b), which shows the PDC difference before and after mutual learning. Generally, T1&T2 have the highest decrease of PDC values after learning followed by T2&T3 and T1&T3. This is another possible reason for the results shown in Fig. 8. For example, in image T1, the PDC difference in T1&T2 is significantly greater than that in T1&T3, which means more difference is used to improve the performance of classifier in T1&T2. Consequently, T1 has higher *F* scores in T1&T2. Similarly, T2&T3 has a lower PDC difference than T1&T2, hence T2 achieves better results in T1&T2 than in T2&T3. For T3, T2&T3 is a better choice because it has a greater difference than that of T2&T3. In addition, the exploited difference during the mutual learning process may directly relate to the success of MSCE rather than the absolute difference. For example, T2&T3 has the highest PDC. Nevertheless, T2 achieves higher *F* scores in T1&T2 (has the largest PDC gap) rather than T2&T3.

A comparison of the best snow cover maps extracted by the proposed method with $N_l = 1$ from the snow cover maps extracted by SVM with $N_l = 400$ is depicted in Fig. 10. The classification agreement is shown in white for the snow pixels and in grey for the snow-free pixels. Inconsistent pixels are depicted in yellow and red. Besides, the *F* scores of the snow cover maps are also included. In the viewpoint of validation results, the best classifier of MSCE trained on 3 labelled samples ($N_l = 1$) achieves better accuracy on T2 and worse results on T1 and T3 than SVM with $N_l = 400$. However, the gap between them is limited (<0.015). This indicates that MSCE helps recover all the information of a large training set by exploiting the information contained in the unlabelled samples. In the aspect of visualization results, the snow cover in T1 and T3 is mainly concentrated in the areas with high

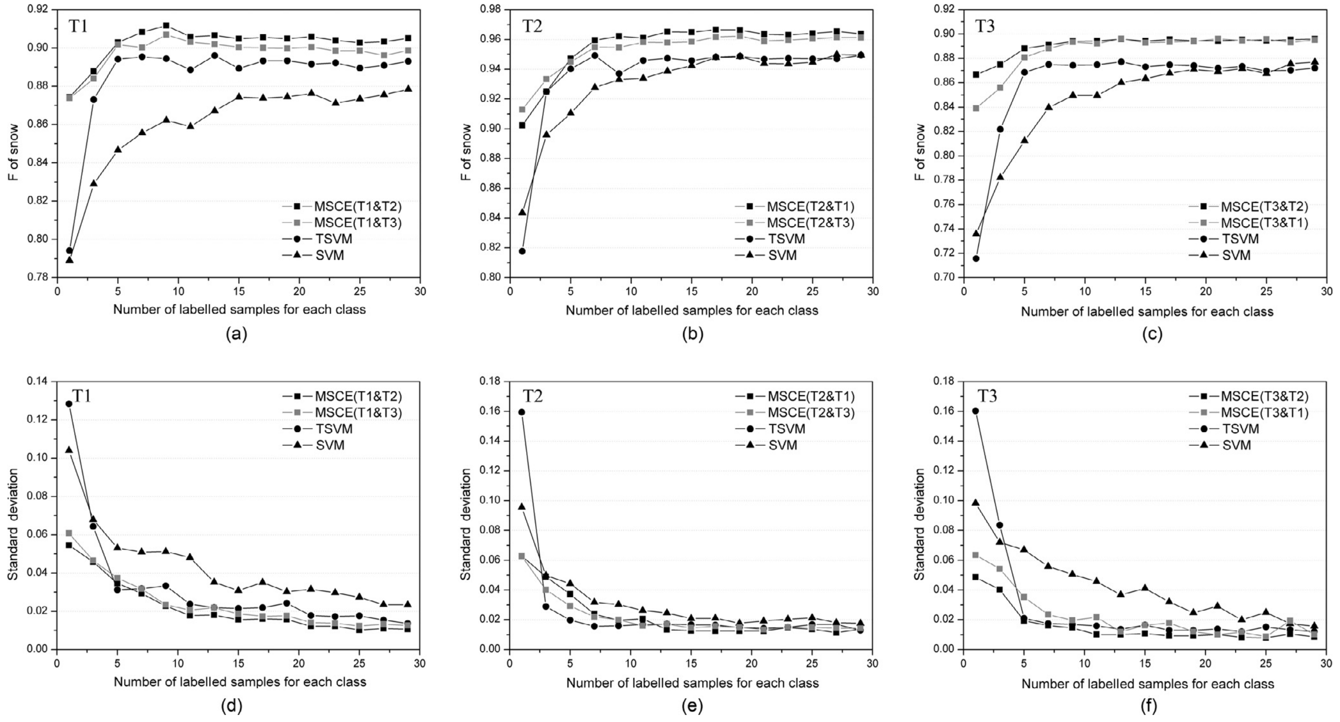


Fig. 7. Average *F* score and standard deviation versus the number of labelled samples N_l. (a), (b), and (c) are the average *F* scores of the three algorithms (SVM, TSVM, and MSCE) on T1, T2, and T3, respectively; (d), (e), and (f) are the corresponding standard deviations.

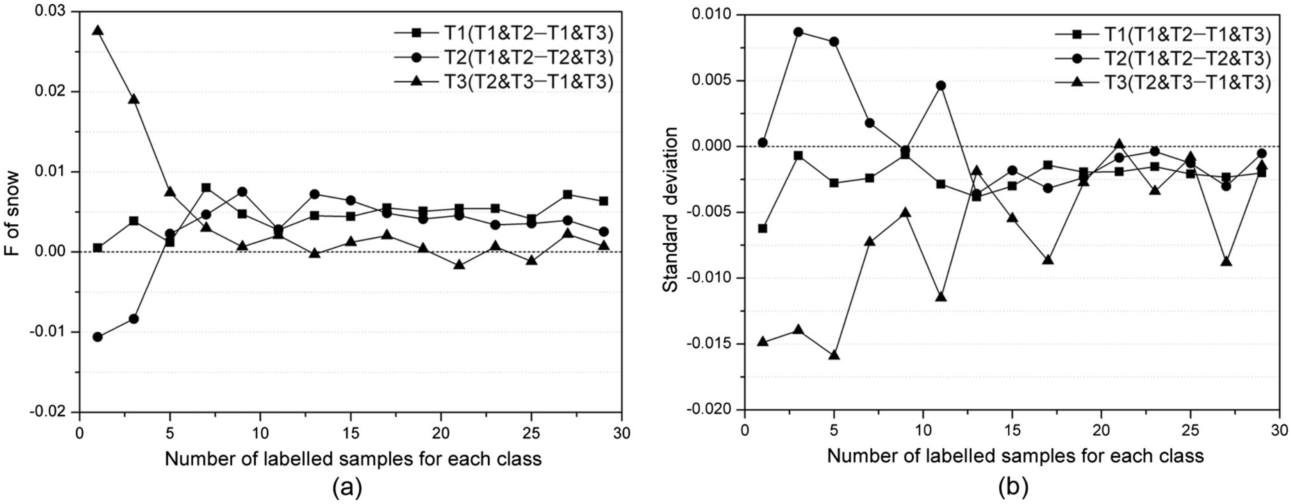


Fig. 8. Impact of different temporal combinations on the extraction results. (a) and (b) respectively show the difference of F score and the standard deviation between different combinations on three images.

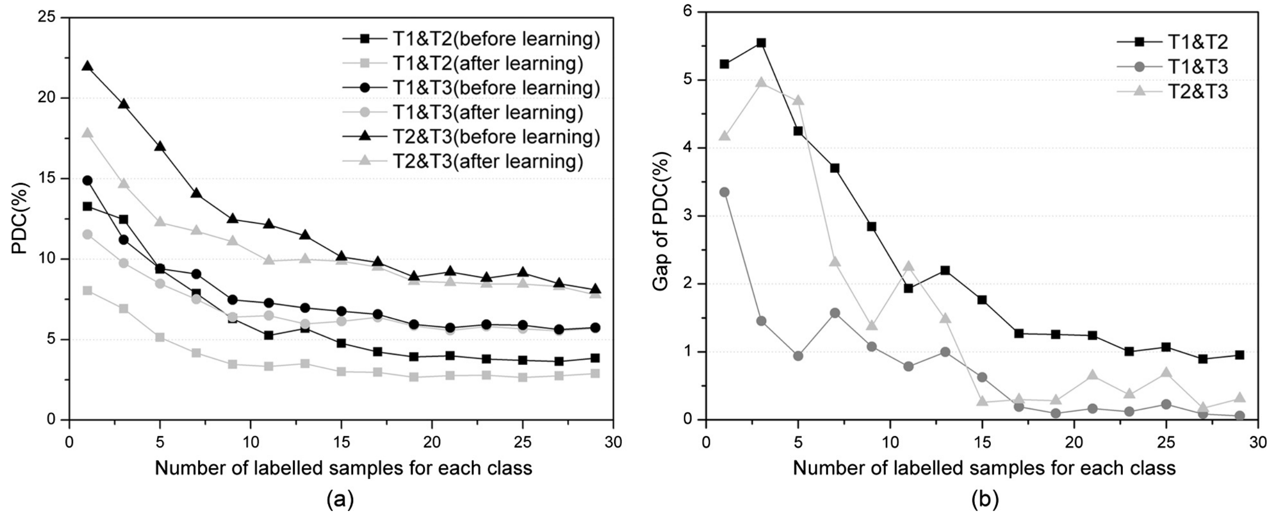


Fig. 9. Prediction differences between classifiers. (a) is the PDC of three temporal combinations before and after mutual learning, and (b) is the difference between PDC before and after mutual learning.

elevation. The area of snow cover in T2 is much greater than that in other images as a result of snowfalls between T1 and T2. Two algorithms result in similar results for T2. In T1 and T3, a large number of pixels are classified as snow by MSCE but interpreted as snow-free by SVM. Most of such inconsistences are located in the area influenced by mountain shadows and the area near the boundaries of snow patches, and presented as small patches.

5.2. Influence of unlabelled sample selection

5.2.1. Influence of N_u

Fig. 11 shows the F scores and SDs based on different N_u and N_l. The highest F score and lowest SD for each N_l are depicted in rectangles. We can clearly see that when N_l increases, N_u that makes MSCE achieve the best performance increases accordingly. That is, the impact of N_u on the accuracy and stability is dependent on the number of available labelled samples. For a small N_l, fewer unlabelled samples result in a higher accuracy and lower SD. However, this phenomenon is reversed when more labelled samples are available, i.e. a larger group of unlabelled samples performs well in terms of both accuracy and stability. Therefore, a small N_u is suggested when very few labelled samples are available, while a large N_u is more suitable when there are adequate labelled samples.

5.2.2. Influence of λ

Experiments on T1 with different λ values were carried out using the combination T1&T2. Fig. 12 shows the F scores based on different λ and N_l. We can clearly find that the impact of λ on the accuracy is quite different when N_l is different. Specifically, where N_l is a small value (see the black line in Fig. 12), the F score gradually increases initially and then decreases as the λ increases. High F scores are achieved when λ has a moderate value (0.8–1.2). Thus, the tradeoff between *confidence* and *difference* is crucial to achieve better results when a few labelled samples are available. In addition, the variation of F scores caused by different λ gradually decreases as N_l increases. This indicates the MSCE based on a smaller group of labelled samples is more sensitive to λ than that trained by a larger set of labelled samples. When N_l is a relatively larger number (see the grey lines in Fig. 12), the F score slightly decreases as λ increases. This means selecting relatively confident

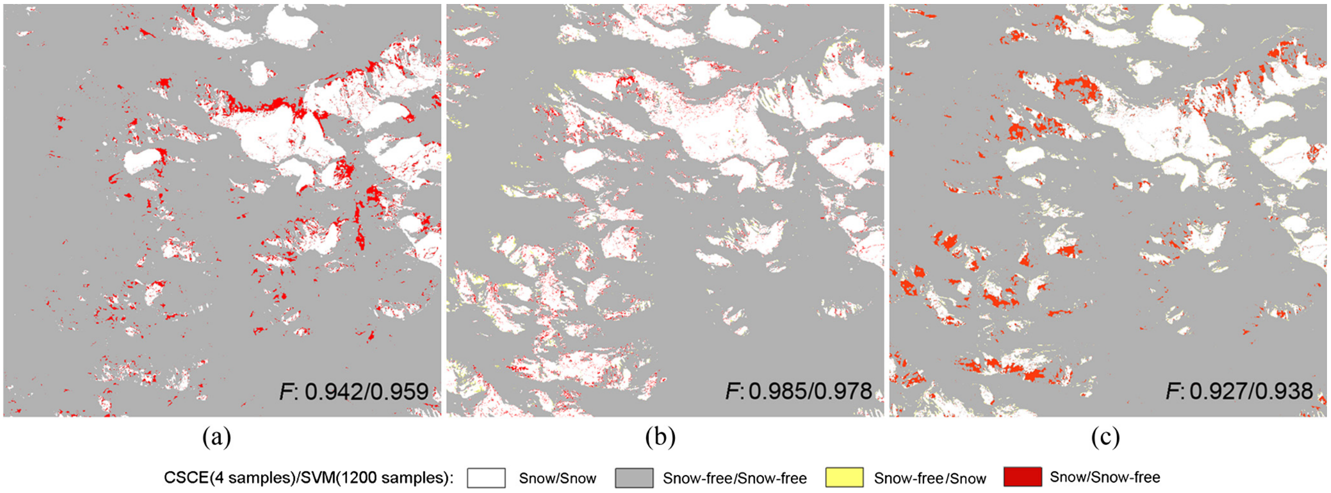


Fig. 10. Comparison of the snow cover extracted by the proposed method with $N_l = 1$ and the snow cover extracted by SVM with $N_l = 400$. (a), (b), and (c) are the results of image T1, T2, and T3, respectively.

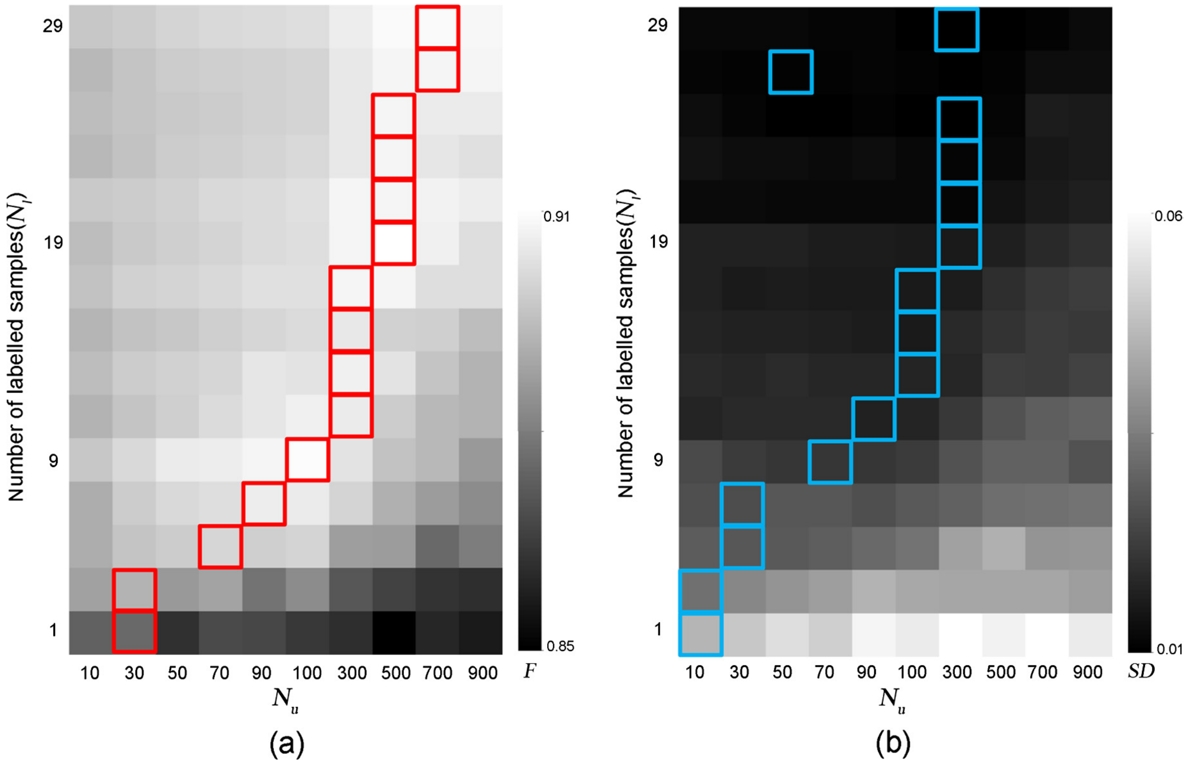


Fig. 11. Impact of different N_u values on the results. (a) represents the *F* scores based on different N_u and N_l (the highest *F* score for each row is outlined by a red square), and (b) represents the corresponding SDs (the lowest SD for each row is outlined by a blue square). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

unlabelled samples is not critical, and it can even have a negative influence on the accuracy. The reasons for these observations are that a small labelled sample set is more likely to train mediocre classifiers under the influence of "bad" information provided by their partners. Hence, the *confidence* and *difference* should be considered simultaneously. On the other hand, the initial classifiers trained by a large set of labelled samples are more robust and less capable of providing "bad" information in the mutual learning process. In these cases, the *difference* is more important. Therefore, a moderate λ ($\sim$1) is suggested when N_l is small, while a small λ is more suitable when adequate labelled samples are available.

5.3. Influence of spatial registration error

Fig. 13(a) shows the *F* scores of T1 versus the simulated spatial registration error in pixels, where a negative value means a bias in the former direction. For example, -10 in the *x*-axis for the north-south means a 10 pixels-offset in north. An interesting phenomenon that the proposed method can benefit from the bias in terms of *F* scores is observed. Larger biases result in higher *F* scores in general. This is especially significant when the bias is greater than 5 pixels. There is no significant difference along different directions of bias in terms of the results. Fig. 13(b) depicts the rela-

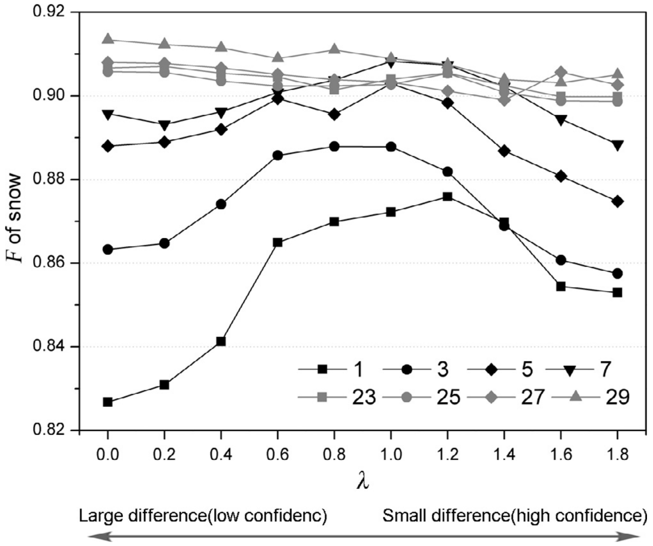


Fig. 12. Impact of different λ values on the results in terms of F score. Different line-symbols indicate the results based on different N_l values.

tionship between the bias and SDs. Obviously, a smaller spatial bias results in a smaller SD when the spatial bias is less than 5 pixels, while these tendencies were reversed as the bias continues to increase. Explanations for those results are that: (i) a change detection procedure is applied to remove the changed area, so that the spatial registration error does not introduce extra uncertainty; and (ii) given a selected unlabelled sample set, the spatial registration error may increase the difference between $p(x^{1*}|y)$ and $p(x^{2*}|y)$, which is the main reason for the success of MSCE. Therefore, we can at least safely make a conclusion that the spatial registration error will not have a negative influence on the proposed method.

6. Discussions

The main principle behind the co-training methods is similar to that of ensemble learning (Zhou and Li, 2010). Different classifiers trained for the same classification task can learn from each other using the labelled samples in ensemble learning, while co-training methods exploits the unlabelled samples in a similar way. In MSCE, two feature subsets of an image are replaced by two images of land surface and thus MSCE can solve multiple classification tasks simultaneously. Similar to the original co-training algorithms, two conditions of MSCE are: (i) each image is sufficient to train a powerful classifier, and (ii) the classifiers trained on two images have a large difference. These conditions have been confirmed by the results observed in Section 5.1. It is worth noting that no pre-procedure is designed in MSCE to ensure these conditions. We assumed that the non-stationary characteristics of the observed reflectance of snow cover in multi-temporal images can provide sufficient difference and the four spectral bands are sufficient to extract the snow cover. However, the difference between two images is not confined to the spectral characteristics and can be further enhanced by texture features. Accordingly, further studies can focus on increasing the difference of classifiers using different feature sets for different images and extracting more powerful feature set for each image simultaneously. A potential approach is to combine feature extraction and selection algorithms with MSCE.

The comparison on the three algorithms indicated that two semi-supervised algorithms (MSCE and TSVM) were superior to the supervised one (SVM) and the mutual learning of MSCE (based on two images) was more robust than the transductive learning of TSVM (based on single image). This is because the mutual-learning procedure of MSCE utilizes the information contained in unlabelled samples in a safer way. If the initial classifier for an image is not powerful, it is difficult for the transductive learning to obtain a better classifier than the initial one (Bruzzone et al., 2006; Vapnik, 1998). On the other hand, the bad initial classifier for an image may be made up by learning from the other classifier in MSCE, which can produce safer results. However, it is still difficult to improve the initial classifiers if both of them are poor. In view of absolute accuracy, MSCE achieved satisfied F scores (∼0.9) when the number of labelled samples per class are greater than 5. This is close to or better than the F scores of binary snow cover maps extracted by other methods from various sources (Graham and Harris, 2003; Rittger et al., 2013; Selkowitz and Forster, 2016).

Despite the satisfied results, more experiments are needed to confirm the effectiveness of MSCE. Although a snowfall followed by a melting process was observed during the acquisitions and it can partly represent the snow accumulation and melting processes in a snow season, more images with different time intervals are needed to test the ability of MSCE in mapping long time series

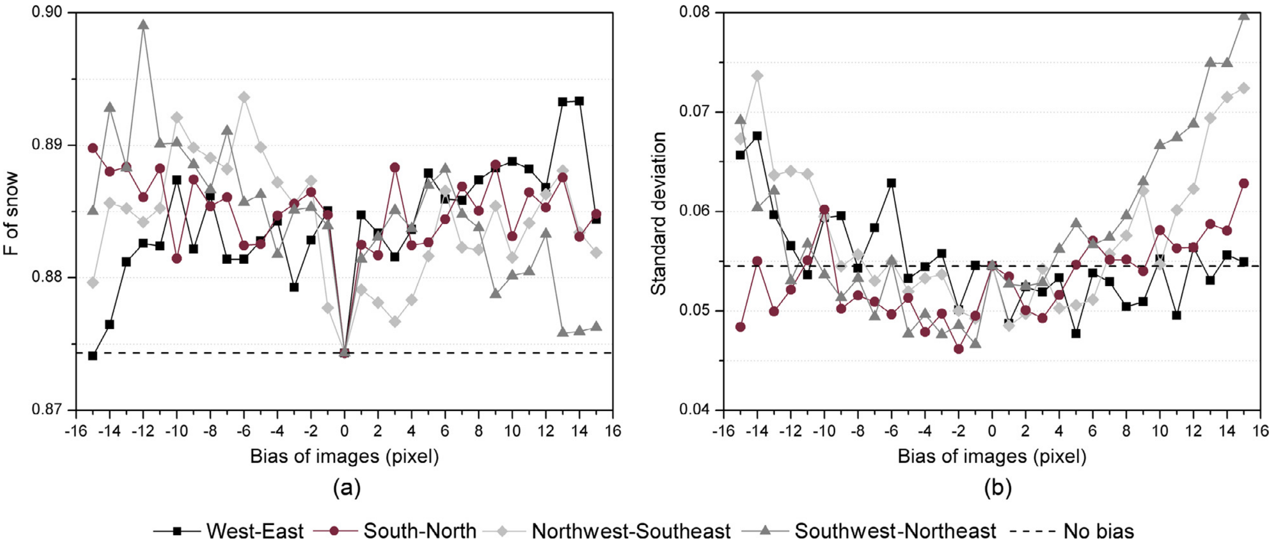


Fig. 13. Influence of spatial registration error. (a) and (b) are respectively the F scores and the standard deviations for snow with different biases in four directions.

snow cover. In addition, the effect of cloud is also not considered. It is worth noting that separating snow from cloud is not straightforward in HSTRRS as shortwave-infrared wavelengths are not available. Current cloud detection algorithms, e.g. the improved Fmask algorithm (Zhu et al., 2015), cannot work. Time series analysis may be more promising because the variation of snow cover happens at a longer interval with respect to cloud, e.g. the method proposed by Hagolle et al. (2010).

In view of practical application, MSCE is more complex than other methods, e.g. SNOMAP (Hall et al., 1995). It at least needs several labelled samples and some user defined parameters including temporal combination, N_u, and λ. However, the results shown in Figs. 11 and 12 provide some suggestions on determining these parameters, which include: (i) selecting a powerful "partner" is essential to achieve better performance, and (ii) a small N_u and a moderate λ ($\sim$1) are more suitable when only a sparse training set is available. Besides, MSCE should be further extended to extract snow cover from more than two images. One promising approach is to replace Co-EM-SVM used in MSCE with other co-training algorithms. Taking tri-training (Zhou and Li, 2005) as an example, it exploits unlabelled samples using three classifiers and thus the combination of tri-training and the framework of MSCE is expected to extract snow cover from three images simultaneously. More generally, inspired by the integration of co-training paradigm and ensemble learning (Zhou, 2011), an ensemble of classifiers trained on time series images can learn from each other using unlabelled samples to derive time series snow cover maps.

7. Conclusions

In this study, a semi-supervised method was designed to map snow cover from HSTRRS. A sparse training set is sufficient to extract the multi-temporal snow cover collectively. Particularly, this method exploits the difference between the observed spectral distributions of two images in a mutual learning way, providing a new strategy to deal with multi-temporal image classification. The proposed method extends the Co-EM-SVM from a single-task method to a multitask one. Such extensions can also be employed in other algorithms of the co-training paradigm.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 41271353, 41671344, and 41601366) and the Qinglan Project of Jiangsu Province (Grant No. 201423). We appreciate thoughtful reviews by anonymous referees.

References

Bahirat, K., Bovolo, F., Bruzzone, L., Chaudhuri, S., 2012. A novel domain adaptation Bayesian classifier for updating land-cover maps with class differences in source and target domains. IEEE Trans. Geosci. Remote Sens. 50, 2810–2826.

Bai, Z., 2013. GF-1 Satellite——The First Satellite of CHEOS. Aerospace China 4, 004.

Bernard, É., Friedt, J.-M., Tolle, F., Griselin, M., Martin, G., Laffly, D., Marlin, C., 2013. Monitoring seasonal snow dynamics using ground based high resolution photography (Austre Lovenbreen, Svalbard, 79 N). ISPRS J. Photogram. Remote Sens. 75, 92–100.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory. ACM, New York, USA, pp. 92–100.

Brefeld, U., Scheffer, T., 2004. Co-EM support vector learning. In: Proceedings of the Twenty-First International Conference on Machine Learning. ACM, New York, USA, p. 16.

Bruzzone, L., Chi, M., Marconcini, M., 2006. A novel transductive SVM for semisupervised classification of remote-sensing images. IEEE Trans. Geosci. Remote Sens. 44, 3363–3373.

Bruzzone, L., Marconcini, M., 2009. Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. IEEE Trans. Geosci. Remote Sens. 47, 1108–1122.

Camps-Valls, G., Marsheva, T.V.B., Zhou, D., 2007. Semi-supervised graph-based hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 45, 3044–3054.

Castelli, V., Cover, T.M., 1996. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. IEEE Trans. Inf. Theory 42, 2102–2117.

Cheng, W., Zhou, C., Liu, H., Zhang, Y., Jiang, Y., Zhang, Y., Yao, Y., 2006. The oasis expansion and eco-environment change over the last 50 years in Manas River Valley, Xinjiang. Sci. China Ser. D 49, 163–175.

Crawford, C.J., Manson, S.M., Bauer, M.E., Hall, D.K., 2013. Multitemporal snow cover mapping in mountainous terrain for Landsat climate data record development. Remote Sens. Environ. 135, 224–233.

D'Addabbo, A., Satalino, G., Pasquariello, G., Blonda, P., 2004. Three different unsupervised methods for change detection: an application. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium. IEEE, Anchorage, AK, USA, pp. 1980–1983.

Dalponte, M., Ene, L.T., Marconcini, M., Gobakken, T., Næsset, E., 2015. Semi-supervised SVM for individual tree crown species classification. ISPRS J. Photogram. Remote Sens. 110, 77–87.

Dasgupta, S., Littman, M.L., McAllester, D., 2002. PAC generalization bounds for co-training. Adv. Neur. Inform. Process. Syst. 1, 375–382.

Dobreva, I.D., Klein, A.G., 2011. Fractional snow cover mapping through artificial neural network analysis of MODIS surface reflectance. Remote Sens. Environ. 115, 3355–3366.

Dozier, J., 1989. Spectral signature of alpine snow cover from the Landsat Thematic Mapper. Remote Sens. Environ. 28, 9–22.

Graham, A., Harris, R., 2003. Extracting biophysical parameters from remotely sensed radar data: a review of the water cloud model. Prog. Phys. Geogr. 27, 217–229.

Hagolle, O., Huc, M., Pascual, D.V., Dedieu, G., 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. Remote Sens. Environ. 114, 1747–1755.

Hall, D.K., Riggs, G.A., Salomonson, V.V., 1995. Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data. Remote Sens. Environ. 54, 127–140.

Hall, D.K., Riggs, G.A., Salomonson, V.V., DiGirolamo, N.E., Bayr, K.J., 2002. MODIS snow-cover products. Remote Sens. Environ. 83, 181–194.

Hinkler, J., Orbaek, J.B., Hansen, B.U., 2003. Detection of spatial, temporal, and spectral surface changes in the Ny-Alesund area 79 degrees N, Svalbard, using a low cost multispectral camera in combination with spectroradiometer measurements. Phys. Chem. Earth 28, 1229–1239.

Hinkler, J., Pedersen, S.B., Rasch, M., Hansen, B.U., 2002. Automatic snow cover monitoring at high temporal and spatial resolution, using images taken by a standard digital camera. Int. J. Remote Sens. 23, 4669–4682.

Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Networks 13, 415–425.

Jackson, Q., Landgrebe, D.A., 2001. An adaptive classifier design for high-dimensional data analysis with a limited training data set. IEEE Trans. Geosci. Remote Sens. 39, 2664–2679.

Joachims, T., 1999. Transductive inference for text classification using support vector machines. In: Proceedings of the 16th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, USA, pp. 200–209.

Jordan, M.I., 1998. Learning in Graphical Models. MIT Press, Cambridge, Massachusetts and London, England.

Kittler, J., Illingworth, J., 1986. Minimum error thresholding. Pattern Recogn. 19, 41–47.

Kurtz, C., Stumpf, A., Malet, J.-P., Gançarski, P., Puissant, A., Passat, N., 2014. Hierarchical extraction of landslides from multiresolution remotely sensed optical images. ISPRS J. Photogram. Remote Sens. 87, 122–136.

Liu, Y., Li, X., 2014. Domain adaptation for land use classification: a spatio-temporal knowledge reusing method. ISPRS J. Photogram. Remote Sens. 98, 133–144.

Matasci, G., Longbotham, N., Pacifici, F., Kanevski, M., Tuia, D., 2015. Understanding angular effects in VHR imagery and their significance for urban land-cover model portability: a study of two multi-angle in-track image sequences. ISPRS J. Photogram. Remote Sens. 107, 99–111.

Negi, H.S., Kulkarni, A.V., Semwal, B.S., 2009. Estimation of snow cover distribution in Beas basin, Indian Himalaya using satellite data and ground measurements. J. Earth Syst. Sci. 118, 525–538.

Nigam, K., Ghani, R., 2000. Analyzing the effectiveness and applicability of co-training. In: Proceedings of the 9th International Conference on Information and Knowledge Management. ACM, New York, USA, pp. 86–93.

Olson, D.L., Delen, D., 2008. Advanced Data Mining Techniques. Springer, Berlin.

Painter, T.H., Rittger, K., McKenzie, C., Slaughter, P., Davis, R.E., Dozier, J., 2009. Retrieval of subpixel snow covered area, grain size, and albedo from MODIS. Remote Sens. Environ. 113, 868–879.

Painter, T.H., Roberts, D.A., Green, R.O., Dozier, J., 1998. The effect of grain size on spectral mixture analysis of snow-covered area from AVIRIS data. Remote Sens. Environ. 65, 320–332.

Riggs, G.A., Hall, D.K., Salomonson, V.V., 1994. A snow index for the Landsat thematic mapper and moderate resolution imaging spectroradiometer. In: Proceedings of International Geoscience and Remote Sensing Symposium IEEE, Pasadena, California, USA, pp. 1942–1944.

Rittger, K., Painter, T.H., Dozier, J., 2013. Assessment of methods for mapping snow cover from MODIS. Adv. Water Resour. 51, 367–380.

Rosenthal, W., Dozier, J., 1996. Automated mapping of montane snow cover at subpixel resolution from the Landsat Thematic Mapper. Water Resour. Res. 32, 115–130.

Salomonson, V.V., Appel, I., 2004. Estimating fractional snow cover from MODIS using the normalized difference snow index. Remote Sens. Environ. 89, 351–360.

Salomonson, V.V., Appel, I., 2006. Development of the Aqua MODIS NDSI fractional snow cover algorithm and validation results. IEEE Trans. Geosci. Remote Sens. 44, 1747–1756.

Selkowitz, D.J., Forster, R.R., 2016. Automated mapping of persistent ice and snow cover across the western US with Landsat. ISPRS J. Photogram. Remote Sens. 117, 126–140.

Shahshahani, B.M., Landgrebe, D.A., 1994. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. IEEE Trans. Geosci. Remote Sens. 32, 1087–1095.

Simpson, J.J., McIntire, T.J., 2001. A recurrent neural network classifier for improved retrievals of areal extent of snow cover. IEEE Trans. Geosci. Remote Sens. 39, 2135–2147.

Sirguey, P., Mathieu, R., Arnaud, Y., 2009. Subpixel monitoring of the seasonal snow cover with MODIS at 250 m spatial resolution in the Southern Alps of New Zealand: Methodology and accuracy assessment. Remote Sens. Environ. 113, 160–181.

Tan, K., Li, E., Du, Q., Du, P., 2014. An efficient semi-supervised classification approach for hyperspectral imagery. ISPRS J. Photogram. Remote Sens. 97, 36–45.

Vapnik, V.N., 1998. Statistical Learning Theory. John Wiley & Sons, New York, USA.

Wang, W., Zhou, Z.-H., 2007. Analyzing co-training style algorithms. In: Proceedings of European Conference on Machine Learning (ECML 2007). Springer, Warsaw, Poland, pp. 454–465.

Warren, S.G., 1982. Optical-properties of snow. Rev. Geophys. 20, 67–89.

Zhang, M.-L., Zhou, Z.-H., 2011. CoTrade: Confident co-training with data editing. IEEE Trans. Syst. Man Cybern. B Cybern. 41, 1612–1626.

Zhou, Z.-H., 2011. When semi-supervised learning meets ensemble learning. Front. Electr. Electron. Eng. China 6, 6–16.

Zhou, Z.-H., Li, M., 2005. Tri-training: Exploiting unlabeled data using three classifiers. IEEE Trans. Knowl. Data Eng. 17, 1529–1541.

Zhou, Z.-H., Li, M., 2010. Semi-supervised learning by disagreement. Knowl. Inf. Syst. 24, 415–439.

Zhu, L., Xiao, P., Feng, X., Zhang, X., Wang, Z., Jiang, L., 2014. Support vector machine-based decision tree for snow cover extraction in mountain areas using high spatial resolution remote sensing image. J. Appl. Remote Sens. 8, 084698.

Zhu, Z., Wang, S., Woodcock, C.E., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. Remote Sens. Environ. 159, 269–277.